Baichuan 2: Open Large-scale Language Models

Year: 2023 | Citations: 902 | Authors: Ai Ming Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian

---

Abstract

Large language models (LLMs) have demonstrated remarkable performance on a variety of natural language tasks based on just a few examples of natural language instructions, reducing the need for extensive feature engineering. However, most powerful LLMs are closed-source or limited in their capability for languages other than English. In this technical report, we present Baichuan 2, a series of large-scale multilingual language models containing 7 billion and 13 billion parameters, trained from scratch, on 2.6 trillion tokens. Baichuan 2 matches or outperforms other open-source models of similar size on public benchmarks like MMLU, CMMLU, GSM8K, and HumanEval. Furthermore, Baichuan 2 excels in vertical domains such as medicine and law. We will release all pre-training model checkpoints to benefit the research community in better understanding the training dynamics of Baichuan 2.